Working Group on Provenance Patterns

Research Data Provenance IG

Last updated Mon, Dec 18, 2017 8:00:08 PM

# Charter

## Overview

Tracking provenance for research data is vital to science and scholarship, providing answers to common questions researchers and institutions pose when sharing and exchanging data.

The tasks for this Working Group focus on finding, detailing and recommending best practices for provenance representation and management.

This group will conduct its work in the manner of a business analysis task: identifying business needs and determining solutions to business problems. Since RDA WGs are not themselves research groups (rather groups of researchers and research agencies), this group will look for existing practice and re-present that for use rather than generate new practice.

The two activity areas of the Working Group will be:

1. Common provenance Use Cases
2. Provenance design patterns

## Deliverables

The deliverables for this Working Group are separated into two time-based cohorts as below. Medium-term goals are mostly about seeking existing practice, and determining possible output forms for the activity areas. Long-term goals are about framing and delivering those outputs.

### Medium-term (M12)

1. A provenance use case recording system[[1]](#footnote-1).
2. An initial collection of provenance use cases, elicited from other interest groups and working groups[[2]](#footnote-2).
3. First documented provenance design patterns generalised from use cases[[3]](#footnote-3).

### Long-term (M18)

1. A taxonomy for provenance use cases.
2. Recommendations for aligning new use cases with provenance design patterns.
3. A sustainability plan for ongoing data and document custodianship.

# Value Proposition

Effective provenance management is sought by many members of the RDA and wider science data community. We propose a working group to help those members adopt existing provenance management practice. This help will be in the form of documenting provenance use cases: centralising a list of them and generalising them to reveal common features and structure.

We propose a working group on *provenance patterns*.

* The patterns should relate to core RDA interests, perhaps data/data and data/people relationships.
* Provenance vocabularies offer a level of generality/specificity that address what we perceive to be implementation gaps.
* Our goal: alerting the RDA community to similarities and common features across use cases in different working groups and interest groups.

### WG activity points

1. **Common provenance Use Cases**
	* Use cases for provenance data or systems are often articulated in terms understood by a particular community however in our group's experience, many provenance use cases are differently worded instances of general cases.
	* The establishment of a published set of UCs would allow people to compare their UCs with known UCs for which recommended implementations and other patterns may already be known. It will also allow people to consider provenance UCs posed by others that may be of future interest to them.
2. **Provenance design patterns**
	* Some ways of doing things in provenance are better than others. This activity is to generate provenance design patterns (for any provenance task such as representation, transmission, use etc.) in response to a series of provenance use cases that we have begun collecting.
	* The patterns should relate to core RDA interests, specifically data/data and data/people relationships.

# Engagement

In addition to serving the RDA community directly, this Working Group aims to serve the immediate interests of existing RDA groups. Provenance is foundational to many other RDA groups' activity and thus maximal impact on the RDA community can be achieved by aligning and assisting work in existing groups. Therefore this working group will engage heavily with other groups and source its primary requirements and exemplars from other groups. RDA Groups with whom we have already communicated are listed in the **Adoption** section. Examples of intersections we believe will be productive over the next ten months include the following:

Reproducibility IG:

The role of provenance models in support of replication.

PID IG:

Requirements for PIDs to maintain provenance content.

Archives and Records Professionals for Research Data IG:

Need for semantic understanding of archived material.

Data Discovery IG:

Upper ontology elements relevant to data discovery.

Preservation e-Infrastructure IG:

Semantic content of preserved data holdings.

Although the following groups are either at or near completion of their work, we see opportunities to respond to portions of their deliverables with our own.

* Publishing Data Workflows WG
* Data Citation WG
* PID Kernel Information WG

# Work Plan

### Timeline

* **Sep 17**: Meeting WGs and summary of activity at Plenary 10.
* **Nov 17 - Apr 18**: Medium-term goals.
* **Mar 18**: Progress report presentation at Plenary 11.
* **May 18 - Oct 18**: Long-term goals.
* **Sep 18**: Draft deliverables at Plenary 12.
* **Mar 19**: Presentation of final reports at Plenary 13.

# Initial Membership

A list of 41 current members[[4]](#footnote-4) is online at <https://www.rd-alliance.org/groups/provenance-patterns-wg>. Since this group's work is highly relevant to other WGs, we will WG members attend this group's meetings either in a liaison role or as members in their own right.

# Adoption

## Other RDA groups

This WG proposal is engagement-driven, initially with other RDA groups, thus it is in them that we expect to see adoption in the short term.

Where another RDA group presents us with a provenance use case, we hope to either:

1. associate that use case with a generic use case and a thus a pre-made generic resolution
2. provide a direct provenance pattern-based resolution directly

In either case, we hope to promote a pattern that the RDA group will adopt and promote to its members.

## Prov WG member institutions

Most current WG members are also members of the Interest Group, and have joined because of they share the WG's goal of identifying and sharing solutions to provenance documentation problems. Therefore outputs from this group, having been generated by its members in response to their direct needs and similar needs of other RDA groups, will therefore be fed back into their home institutions for adoption there.

RDA groups sharing members with this WG with whom we have already discussed collaboration plans are:

* **Storage Service Definitions WG** - Nicholas Car
	+ storage definition ontology to use PPWG patterns
* **RDA/TDWG Metadata Standards for attribution of physical and digital collections stewardship** - Andrew Dormer
	+ Entity/Agent relations to use PPWG patterns
* **Agrisemantics WG** - Andres Ferreyra & Simon Cox
	+ ADAPT model being mapped to PROV by PPWG
* **Data Versioning WG** - Jens Klump & Lesley Wyborn
	+ DV WG UCs being placed into the Patterns DB
* **Physical Samples & Collections IG** - Andrea Thomer, Nicholas Car, Jens Klump, Lesley Wyborn
	+ Will adopt PPWG patterns for both digital & physical entity provenance
* **Vocabulary Services Interest Group** - Simon Cox & Nicholas Car
	+ Will adopt vocab term provenance from PPWG patterns

## Non-RDA groups

The international provenance research community is in contact with many potential consumers of provenance patterns due to their profile as experts on provenance. The potential consumers don't always receive the advice they are seeking due to differences in their aims and that of the research community's. The research community needs to push the provenance envelope forward and not dwell on previous work, even when that work may contain patterns perfectly suited to the potential consumers' needs.

The Provenance IG has and the WG will have good contacts with the international provenance research community with several IG members having made substantial contributions to provenance research initiatives such as the [Open Provenance Model](http://openprovenance.org/), [ProvONE](http://vcvcomputing.com/provone/provone.html), the [PROV W3C standard](https://www.w3.org/TR/prov-overview/) and having presented at many recent provenance conferences such as [IPAW 2014](http://www.ipaw.info/ipaw14) & [2016](http://www.ipaw.info/ipaw16), [TaPP 2015](http://workshops.inf.ed.ac.uk/tapp2015/), [TaPP 2017](http://batesa.web.engr.illinois.edu/tapp17/).

The full listing of the IG's involvement in provenance conferences is available on the RDA Prov IG wiki:

* <https://www.rd-alliance.org/group/research-data-provenance/wiki/provenance-conferences>

Formal links to the following non-RDA provenance groups will be maintained through WG and other group joint members. Relations and actions already established are:

* **W3C's Data Exchange WG** - Nicholas Car & Rob Atkinson
	+ harmonising of provenance Use Cases
* **ISO TC211** - Simon Cox & Ivana Ivanova
	+ patterns for associating geographic metadata with provenance under review by TC211 members
* **ENVIRI Plus** - Barbara Magagna & Markus Stocker
	+ ENVRI UCs to be represented in UC DB
* **Australian Government Linked Data WG** - Nicholas Car, Simon Cox & Ivana Ivanova
	+ Recommendations to Aust. government for provenance will be to use PP WG patterns
* **DARE - Delivering Agile Research Excellence on European e-infrastructures** - Alessandro Spinuso
	+ e-infrastructure partners will be asked to record provenance Use Cases in UC DB
* **Provenance Week organising committee** - Paolo Missier
	+ Plans for post PP WG provenance tool maintenance will be workshopped
* **International GeoSample Number Technical Committee** - Nicholas Car & Jens Klump
	+ Mapping of IGSN metadata to PROV already taking place

## Continued adoption

Some of the outputs from this WG are targeted at continued adoption over time, such as the best practice documents (Use Cases and their corresponding Patterns). This proposal includes a deliverable for a "sustainability plan for ongoing data and document custodianship" after having initially established a database of use cases and provenance patterns. Such a plan is currently missing from the international provenance community despite widespread recognition that it would be useful. This was recognised at IPAW 2016, independently of any RDA involvement and will be discussed at IPAW 2018 by the provenance research community and several attending WG members.

It is expected that at the conclusion of this WG, the current provenance IG will have some role in the custodianship of its outputs.

1. Our use case and patterns database has been online since June 2017: <http://patterns.promsns.org/>. [↑](#footnote-ref-1)
2. Thirteen entries are online as of December 2017: <http://patterns.promsns.org/usecases>. [↑](#footnote-ref-2)
3. Ten pattern records are online as of December 2017: <http://patterns.promsns.org/patterns>. [↑](#footnote-ref-3)
4. December 2017. [↑](#footnote-ref-4)